Nebraska Department of Education, March, 2006

The Process for…..

A GUIDE FOR ASSURING THE TECHNICAL QUALITY OF CLASSROOM ASSESSMENT

Introduction

The purpose of this document is to provide easy, clear directions for classroom teachers or others for applying the Six Quality Assessment Criteria to classroom assessment. Although the directions in this document apply to all assessment, the methods selected are most appropriate for performance-based classroom assessment.

Please Note:
If these criteria are applied to classroom assessment, the results of those assessments may be used for state reporting.

THE SIX QUALITY ASSESSMENT CRITERIA

1. **The assessments reflect the state/local standards.**
2. **Students have the opportunity to learn.**
3. **The assessments are free of bias and insensitive situations.**
4. **The assessments are at the appropriate level.**
5. **The assessments are reliably scored.**
6. **The assessment mastery levels are appropriately set.**

Who Should Do This Work?

Panels of experienced and veteran teachers are qualified to do this work. A leader should be designated and all steps should be documented.

Important Note: Even though each of the Criterion is identifying a key individual element, the criteria are interrelated and all six work as a system. For example: The process for establishing Performance Level Definitions for Advanced, Proficient, Progressing and Beginning is described in Criterion Six (page 11,) but the performance level definitions are needed in Criterion One, and when using the Teacher Judgment Decision Consistency model in Criterion Five.

Criterion One

The assessment reflects the state/local standards.

This criterion assures match to standards (validity) and sufficiency (adequacy of sufficient opportunity to demonstrate skill or knowledge.)

1. Convene a group (3-7) of experienced professional educators familiar with the content and grade level being assessed. It is recommended that several grade levels should be represented (e.g. If a fourth grade assessment is being examined, include experienced teachers from grade 2-5.) **Note: This panel must be independent reviewers. They may not be the assessment writers or developers.**

2. The panelists should examine the standard and the assessment task/item(s) and make an independent decision about whether they match in content and in complexity.

3. Each panelist should make independent decisions about whether or not the assessment task/item(s) capture the essence or main purpose of the standard.

4. After the independent decisions are made about match to standards they should be recorded on a sheet (Worksheet A.) Panelists should identify the name of the assessment and the item type (subjectively or objectively scored).

5. The panelists should decide how consensus will be reached (100% agreement, talk through until the majority agree, 6/7 agree, etc.)

6. Explain the task of sufficiency to the group. Each individual will be making independent decisions about the difficulty level of each assessment task/item. Does the assessment include enough tasks/items so that students at all levels (beginning, progressing, proficient and advanced) can demonstrate their skill/knowledge? Individual decisions should be recorded on the worksheet (Worksheet A.)

7. The independent group needs to know the performance level definitions that have been developed with the assessments (see Criterion Six) in order to determine sufficiency. Performance level definitions are descriptors of student performance at advanced, proficient, progressing, and beginning levels.

8. Decide how the group will come to consensus agreement on decisions of sufficiency. Determine the decision rules.

9. Determine and record final decisions about sufficiency and any needed changes to be made on a single record sheet, but keep the individual panelists' worksheets as documentation.

10. Develop a plan and a timeline to make and review any of the needed changes in the assessments.

The students have the opportunity to learn.

This criterion assures that the standards are present in the local curriculum and that students have been taught at least 80% of the content prior to being assessed on it.

1. Convene a group of educators who teach the local curriculum. This needs to be done by each individual district. For collaborations and consortiums, therefore, this needs to be conducted and recorded for each district. Unlike the panel used in Criterion One, these educators do not need to be independent. They may include the assessment developers. This group should consist of the teachers who teach the local curriculum and who give the assessments.

2. Panelists should examine the local curriculum guide and other relevant material to identify which standards are taught in which unit, and at what time during the year.

3. Once the independent decisions are made, the group should talk through the decisions about when standards are taught until coming to agreement. They should record all decisions on Worksheet B.

4. Panelists need to agree as a group when the assessments should be given in relationship to instruction so that students have the opportunity to receive instruction on 80% or more of the standards prior to assessment. Those dates (or approximate times during the year) need to be identified and recorded.

5 Any redundancy of standards or absence of standards needs to be identified and noted. The same is true for any inappropriate timing of instruction or assessment.

6. A plan and a timeline for addressing any needed changes in opportunity to learn needs to be developed and the appropriate changes made.

Criterion Three

The assessment is free from bias and insensitive situations.

This criterion assures that a reviewer panel has examined the assessment for fairness and effectiveness, that nothing in the assessment or the directions is inappropriate, unkind, demeaning, or unclear.

1. The bias review is best conducted by a panel of people who were not assessment developers. This is a recommendation for Criterion Three but not a requirement as it is in Criterion One.

2. A qualified leader should conduct training in assessment bias for all who will be reviewing for bias, using examples of what is considered as bias including unfair penalization and offensiveness.

3. The reviewers should practice identifying examples of unfairness and offensiveness on sample assessments.

4. The panel members should then independently examine the assessments used, identifying any possible instances of unfairness or offensiveness. They should record their responses for each item on Worksheet C.

5. The panelists should then determine collectively the instances of unfair penalization and offensiveness needing to be changed in the assessments. Needed changes need to be documented.

6. The final decisions of the group should be recorded on Worksheet C and a plan and timeline for making the needed changes should be developed.

Criterion Four

The assessment is at the appropriate level.

This criterion assures that the cognitive (thinking) level of the assessment is appropriate for the grade level being assessed.

1. A panel of educators familiar with the grade level and content should review the assessment for appropriate level. It is recommended that a span of grades be represented (e.g. 3-6 grade teachers for a 4th grade assessment.) It is helpful to include special education teachers, a school psychologist, and a counselor in the group.

2. The panel should review the assessments and make a decision about appropriate level. They should review the nature and content of the tasks and determine whether the assessment approach is appropriate for that grade level.

3. The panelists should talk through their decisions as a group and determine how they will come to consensus.

4. Any needed changes and recommendations should be noted along with the final decisions about appropriate level. All should be recorded on one final Worksheet D.

5. A plan and timeline should be developed for making the needed changes in the level of the assessments.

Criterion Five

The assessments are reliably scored.

This criterion assures the reliability and consistency of scores so that educators can have confidence in the student performance results generated by the assessments.

The reliability values are calculated as an average percentage across all standards and must meet or exceed .70 to be considered acceptable.

The method of calculating reliability is determined by the type of assessment (objectively or subjectively scored) and the number of students assessed.

Method	Type of Assessment	Number of Students Assessed
Internal Consistency (KR-20), KR21, Coefficient Alpha, Split Half)	Objectively Scored	May need large number of students for stable results (30 or more)
Decision Consistency, Test-retest, Parallel Forms	Objectively Scored or Subjectively Scored (if the two decisions are independent	May be used with any number of students
Inter-rater Reliability	Subjectively Scored	May be used with any number of students

Internal Consistency Methods (KR20, KR21, Coefficient Alpha, Split Half)

1. These methods are most appropriate only for groups of 30 or more students and for objectively-scored assessments. Small schools could collect results over multiple years to reach 30 or join with other districts to reach sufficient numbers.

2. These methods are most easily computed using computer software. They involve entering data results into a program and generating percentage values.

3. The directions for each statistical analysis program must be learned and followed.

4. If assessments are administered to the same student at multiple times, the results of the first administration should be used in the internal consistency calculations.

5. This method does not rely on teacher's professional judgment in the calculation.

Decision Consistency Methods - Primarily used for objectively scored assessment.

1. This method is used primarily with objectively scored assessments, but may be used with subjectively scored assessments (if the two decisions about students' performance are independent.)

2. This method is helpful to small districts but can be used with any number of students.

3. This method requires two independent decisions about student performance. Basically, this method involves the calculation of the percentage of times the two decisions agree. The two decisions could be based upon any of the following:

    - Assessment results from two assessments measuring the same thing at the same level of difficulty. Both assessments would have to meet the Six Quality Criteria.

    - Assessment results from CRT and results from an NRT (again, the CRT would need to have been run through the Six Quality Criteria.) This approach could only be used with those standards that have been determined to match the NRT's.

    - Teacher judgment and assessment results.

The last method, Teacher Judgment and Assessment Results can be calculated. It will be called – Teacher Judgment Decision Consistency.

Teacher Judgment and Decision Consistency

a) Teachers participating in this reliability method need to review the Performance Level definitions that were developed in Criterion Six and used by the independent review team in Criterion One to examine the assessment sufficiency.

Through this review the teachers will all have the same understanding of student performance at each of the levels: advanced, proficient, progressing, and beginning. During training, some time needs to be spent discussing the performance level descriptors so that common understanding occurs.

b) Based on the performance level definitions, teachers make an independent professional decision about the performance level they believe their students will achieve. This judgment needs to be made before the teachers know the assessment results.

The teachers' judgments are recorded on Worksheet E. They may be made by standard, by groups of standards (strands), or by assessments.

c) The actual calculations of reliability cannot be completed until the assessments are scored and mastery levels determined.

d) Once the scoring is done and mastery levels set, the rest of the worksheet can be completed.

The results of the actual assessments are recorded by actual mastery level achieved. If the teacher judgment and the actual results are identical " + " is recorded in the column as agreement; if not, the match is recorded as "0" in the agreement column.

5. Convert the total number of decisions that agree into a percentage. The percentage across all standards, strands, or assessments may be arranged for a total reliability calculation.

Inter-rater Reliability Used for subjectively scored assessment.

The decision consistency method described earlier calculates the agreement between two independent decisions. The inter-rater reliability, on the other hand calculates a decision between two independent raters.

1. Subjectively scored assessments are scored with a rubric or clearly written criteria outlining specific expectations for assessment results.

   The raters must be thoroughly trained on the rubric and must be clear about the expectations of the assessment. If the rubric has fewer than 6 score points only exact match agreement may be calculated.

2. Examples of the assessment results (products or anchor papers) at all mastery levels need to be shared with the raters during the scorer training so that the raters know what the assessment product results look like at all levels.

3. Raters score the assessments independently and record their scores independently.

4. The rater agreement is calculated by determining how frequently the independent judgments of the raters agree about the level of performance on the assessment. A process needs to be in place in case the two raters do not agree.

5. The final number of exact agreements are calculated and converted into a percentage.

   $$\frac{\text{Number of exact agreements}}{\text{Number of possible agreements}}$$

   The overall reliability is calculated by averaging the reliability across all standards, all strands, or all assessments.

The assessment mastery levels are appropriately set.

This criterion assures that mastery levels have not been set arbitrarily, and that everyone has come to agreement on what the mastery levels (advanced, proficient, progressing, and beginning) mean.

Mastery levels are appropriately set when three things are integrated in a process: agreed-upon performance level definitions, professional judgment, and actual student results.

The Establishment of Performance Level Definitions

1. A group of teachers familiar with the content and the grade level students being assessed can develop these definitions. These teachers may be those who wrote or administered the assessments.

   **Note:* These performance level definitions are the same set used in Criterion One to review for sufficiency and the same set that are used in Criterion 5 if the Decision-Consistency and Teacher Judgment method is used to calculate reliability.

2. To establish performance level definitions, the leader initiates a discussion with the group about the characteristics of the "barely advanced" students' performance. Those characteristics are put on the board or poster paper and placed in front of the group. The same conversation and recording of characteristics occur about the "barely proficient" and the "barely progressing" students. The group discusses the differences between each group.

3. From the characteristics of each category definitions are drafted for each of the four categories: advanced, proficient, progressing, and beginning student performance. Through this process of building the definitions together, the entire group comes to consensus about what these definitions mean. Everyone agrees on the common language.

Next, the decision must be made whether to use a student-based method or a test-based method.

Student-based Method (Modified Contrasting Group Method)

Note: Student-based methods are typically inappropriate in small schools unless multiple years of data are being calculated.

- Panelists must know the assessed students.
- Not always reliable with fewer than 30 students being assessed

Test-based Method (Modified Angoff Method)

Note: Panelists need to have content knowledge and familiarity with students at the appropriate grade.

- Panelists may or may not know the assessed students.
- May be used with any number of assessed students.

Modified Contrasting Group Method (Student-based Method)

a) The panelists (who know the assessed students) review the performance level definitions and become familiar with them, talking through the understanding of what each definition means.

b) Prior to knowing the assessment results and based on the knowledge of the student, each panelist makes a professional judgment decision at which level each student will score. An "X" is placed in the blank of the predicted performance level. Those decisions are recorded on Worksheet F.

c) After the scores of the assessment have been completed and totaled, the student results replace the professional judgment decision, so that the "X" is replaced by student results.

d) Once results have replaced the professional judgment "X's", the columns of numbers are averaged vertically. That results in four columns of numbers averaged.

e) Lastly, the adjacent columns of averaged numbers are averaged so that the average is taken of the beginning and progressing column, progressing and proficient column, and the proficient and advanced column.

f) The new averages become the "cut scores" and the results are ranges for advanced, proficient, progressing, and beginning.

Modified Angoff Method (Test-based Method for Objectively Scored Items)

Teachers using this method must know both assessment content and the characteristics of the students taking the assessment. Teachers will analyze each **item** on the assessment in relationship to student performance. It is not sample dependent; the size of sample does not generally influence results.

a) Teachers identify in their minds a student who barely achieves at each proficiency level. (Refer to the Proficiency Level Definitions.)

b) If at the point in this process you determine that the student would get an item correct, write an "R" on the line for that level and each proficiency level above that point.

c) Consider the barely progressing student. If you would expect them to answer item 1 correctly, put an "R" on the line.

d) Then consider the barely proficient student. If you would expect them to answer item 1 correctly put an "R" on the line.

e) Look at item number 1 again. Would you expect the barely advanced student to answer this question correctly? If so, put an "R" on the line.

f) Then consider the beginning student. If you would expect them to answer item 1 correctly, put an "R" on the line.

g) Continue this method for each item on the assessment.

h) To compare the cut scores add up the number of "R's" for each performance level. Record at the bottom of each column.

i) Use the number of "R's" to determine your minimum cut score for each level of proficiency and to set mastery ranges.

j) Record the range for each proficiency level in the boxes provided at the bottom of your chart. These are your mastery levels.

<u>Modified Analytical Judgment</u> (Test-based Method for Subjectively Scored Tasks)

You will need anchor papers or exemplars.

a) Discuss each student: Barely Advanced, Barely Proficient and Barely Progressing. Discuss what their work will look like.

b) Select multiple exemplars for each score point (they need to be scored ahead of time but the teachers do not know the scores) for each level.

c) Have each panelist separate papers into three categories: below proficient, proficient, above proficient.

d) Have each panelist find the three best papers from the group classified as below proficient.

e) Have each panelist find the three poorest papers from the group classified as being proficient.

f) For the six papers selected, take the average of the actual scores.

g) Calculate the average across panelists – average the averages. The answer becomes the final cut score.

MATCH TO STANDARD AND SUFFICIENCY

Standard #	Assessment Name	Match		# of Tasks/ Items	Beg.	Prog.	Prof.	Adv	Changes Needed
		Yes	No						

OPPORTUNITY TO LEARN

Standard	Dates Unit Taught	Assessment Used	Dates Assessed	Changes Needed

ASSESSMENT HAS BEEN REVIEWED FOR BIAS

Training Date(s) :		
Assessments	**Dates/Review**	**Changes Made**

LEVEL IS APPROPRIATE FOR STUDENTS

Assessments	Appropriate Level		Recommendations
	Yes	No	

TEACHER JUDGMENT

Student Name	Tchr Judge	Results	Agree	Tchr Judge	Results	Agree	Tchr Judge	Results	Agree	Tchr Judge	Results	Agree	Tchr Judge	Results	Agree	Total Agree

MODIFIED CONTRASTING GROUP METHOD

Student Name	Beginning	Progressing	Proficient	Advanced

MODIFIED ANGOFF METHOD

Standard: _____ Date Calculated: _____

Assessment Title: _____ Level _____

Item #	Barely Progressing	Barely Proficient	Barely Advanced
1.			
2.			
3.			
4.			
5.			
6.			
7.			
8.			
9.			
10.			
Total			

Advanced	Proficient	Progressing	Beginning